Extracting roof sub-components from orthophotos using deep-learning -based semantic segmentation

Jiajun Li[1], Boan Tao[1], Frédéric Bosché[1], Chris Xiaoxuan Lu[2] and Lyn Wilson[3]

[1]School of Engineering, University of Edinburgh, UK
[2]School of Informatics, University of Edinburgh, UK
[3]Historic Environment Scotland, UK

Jiajun.Li@ed.ac.uk, boan.tao@ed.ac.uk, f.bosche@ed.ac.uk, xiaoxuan.lu@ed.ac.uk, lyn.wilson@hes.scot

Abstract -

Best practice for the detection and annotation of visible defects in slated roofs is by annotation of photos, ideally orthophotos. If such a process is to be effectively automated in support of emerging Digital Twinning solutions, it is necessary to first recognise the external sub-components of the roof in the orthophotos, in particular the slated and leadwork areas. Using a dataset composed of many photos from two historic buildings, this study develops and compares different deep-learning -based semantic segmentation models to segment roof orthophotos into slated areas, leadwork, and 'other' areas. Since orthophotos typically contain pixels which do not belong to the roof panel (black 'background' pixels), the method employs a subsequent 'background' label correction step. The best-performing model is found to be PointRend with Focal Loss: overall *aAcc* = 99, *mIoU* = 88.91, and *mAcc* = 92.77; for slate class, *IoU* and *Acc* is nearly 100; for leadwork class, *IoU* and *Acc* is around 90.

Keywords -

Semantic segmentation; Deep learning; Slated roof; Orthophoto

1 Introduction

For most traditional slated building roofs, they are mainly composed of different elements: slate roofing tiles, leadwork, stonework, and masonry [1]. Additional decorative or functional accessory components may also exist, e.g. ventilator, balustrade, satellite, etc. Slates occupy the largest area and are where deteriorations happen easily and frequently, especially due to weathering with time [2]. To better detect the visible defects in roof monitoring, the annotation of photos is a common practice. The best practice to date employs orthophotos due to their benefits for length and area quantification.

Within the emerging area of Digital Twinning in the built environment [3], there is great interest in automating data acquisition and processing for building fabric monitoring, in order to efficiently, robustly and safely detect and monitor defects, and support computer-aided maintenance decision making. In the context of slated roofs, we showed in [4] how orthophotos can be generated for individual roof panels, from UAV-acquired photogrammetric data and the building's 3D digital twin model. To ensure effective defect detection in each such orthophoto, it is next necessary to distinguish the sub-components inside the orthophoto, in particular the slated and leadwork areas.

Semantic segmentation was developed decades ago, and can be applied to different kinds of data, from 2D image to 3D point cloud, and even video data. When applied to 2D images, it enables assigning a class label to each pixel of an image, and such pixel-level semantic information can help make judgements or be leveraged by other tasks [5, 6]. With the development of deep learning, different deep architectures have been introduced, especially Conventional Neural Networks (CNNs). As a result, the performance of semantic segmentation models has been greatly improved, not only in accuracy but also in efficiency [7].

In construction informatics research dealing with 2D and 3D data, such as in Scan-to-BIM, semantic segmentation is increasingly regarded as an essential step after data collection, to provide further information useful to subsequent tasks such as object detection [8]. This technique has been applied to different types of data, including: (1) 2D image of indoor scene [9] and aerial images of different architecture [10]; (2) 3D point cloud of building interiors [8], plumbing and structural components [11], autonomous vehicles and robot navigation [12].

This paper reports on the development and comparison of well-established deep-learning -based semantic segmentation models for segmenting orthophotos of individual roof panels into 'background', 'slate', 'lead', and 'other' classes.

2 Related work

With its powerful pixel-level segmentation ability, semantic segmentation has been developed into a well-established tool in Computer Vision.

Most recently, different deep learning models have been proposed for 2D image data. *DeepLabV3*, a widely used

system developed by Google [13], can handle the problem of segmenting objects at multiple scales with cascaded module and Atrous Spatial Pyramid Pooling(ASPP). Compared to *DeepLabV3*, *PointRend* demonstrates better performance (higher mIoU value), by extracting the point features made up on fine-grained features and coarse prediction [14]. Specifically, *PointRend* handles better the problems of smooth region and blurry contours from bilinear sampling process. Thanks to the recognisable features, the object can be easily detected and classified for this task, but the contour detection should be more precise.

In recent years, Vision Transformer (ViT) has achieved superior performance to the commonly used CNNs, by splitting each image into patches. Due to their outstanding performance in image classification, these models have also been explored for semantic segmentation: *Segmenter* can capture global interaction between elements of a scene using transformer, rather than the traditional convolutions, which would easily cause information loss [15]. *SegFormer* reaches better performance and efficiency by redesigning the Transformer encoder and uses a simple multilayer perceptron (MLP) decoder[16]. However, all authors also point out that ViT relies heavily on large-scale datasets to achieve good performances [17].

Within these main classes of deep learning models, different variants can be created by modifying different some components, such as the loss function. Focal Loss is a commonly used loss function for dealing with class imbalance and putting more focus on the hard and misclassified examples, by multiplying each class loss with a weighting factor [18]. Dice Loss can deal with the imbalance problem between foreground and background, by giving more importance on foreground than background, thus making the model region-related [19]. These two loss functions and the default Cross-Entropy Loss can be paired and used together in improving model training [20].

3 Method

We assume as input an orthophoto of a slated roof panel generated by the method presented in [4]. The generated orthophoto shows the roof panel, which is composed of different essential sub-components, slates and leadwork, as well as other less frequent elements (e.g. stone, grid, equipment, glass, ladder, etc.), as can be seen in Figure 5. Since the intention of this work is to find defects in slated areas as well as leadwork areas, we must first segment the orthophoto to isolate these particular subparts. For this, we explore different methods for (pixel-level) semantic segmentation.

3.1 Dataset

The data output of Duff House in Banff, Scotland by Li et al. [4] includes data for 36 panels. We split these into 27 panels for training and 9 panels for testing. For each panel, we retain maximum 10 unmerged orthophotos obtained from different photos of the same panel (with different angles and covering the panel in various ways). This dataset is used for developing and comparing different initial semantic segmentation models.

In addition, the data output of St Mary's Church, in Stirling, Scotland includes data for 18 panels. With this data, we created another dataset composed of one unmerged orthophoto per panel. This dataset is used solely to test the generalisability of the models created using the Duff House dataset.

Finally, a combined dataset using data from the two buildings is created. It is divided into a training dataset that is ~80% of the overall dataset (27 panel orthophoto data of Duff House, along with 13 panel orthophoto data of St Mary's Church) and a testing dataset that is the remaining ~20% (9 from Duff House, along with 5 from St Mary's Church). Here, only one unmerged orthophoto is kept for each panel. We select the orthophoto with the largest coverage of the panel.

All the orthophotos used in the datasets above are manually labelled in 4 classes: *background* (labelled as 0), *slate* (1), *leadwork* (2), and *other* (3). The *other* category contains: stone, grid, equipment, glass, ladder, etc. The class *other* normally occupies a very small proportion of pixels in orthophotos.

3.2 Deep learning model

As discussed earlier in 2, these are the state-of-art models that can be useful to the specific problem in this study: *DeepLabV3*, *PointRend*, *Segmenter*, *Segformer*. Basic hyper-parameter settings are tuned for best performance by comparing these deep learning models. Our settings are reported in Table 1. All models were pre-trained using their default weights.

Table 1. Experimental parameters

Parameter	Value
Batch size	4
Max iteration	2400
Validation interval	400
Training vs. Validation	75% : 25%

During training, the input images are not rescaled. This is because most of the roof orthophotos are rectangular with varying width-to-length ratios, and the resizing processing operations that are typically applied in semantic segmentation pipelines would result in information loss. But, to meet the requirements of input image size and com-

puting capacity limits, all images are cropped (tiled) and read as $512 \times 1024 \times 3$ matrices.

As will be shown in the Experimental Validation (Section 4), PointRend with default backbone (ResNet) and loss function (Cross-Entropy Loss) achieves best performance all four initial models. As a result, further experiments are conducted by using different loss functions including Dice Loss (sensitive to region detection) and Focal Loss (sensitive to imbalance problem). With grid search as a tool for hyperparameter adjustment, specific weights between different losses are selected for best performance.

The models trained with the Duff House dataset were then tested with the St Mary's Church dataset to assess its generalisation ability. The results lead us to finally use the combined orthophoto dataset (Duff House + St Mary's Church) for training and testing a final model with best segmentation performance and generalisation ability.

3.3 Background label correction

In most deep learning semantic segmentation methods, precise delineation of segmentation boundary is a challenging problem, with many confusions arising at those boundaries. Figure 1 illustrates this issue with a typical output of the semantic segmentation models we explored. The figure highlights the discrepancies between the *Prediction* and *Ground Truth* (*GT*) for the different classes.

However, in this study, the set of *background* pixels is actually known a priori, because the panel boundary is defined by the Digital Twin 3D model projection (see [4] for details). In other words, the Ground Truth for the *background* class is known a priori. Therefore, an extra step is introduced to correct the False Positive (FP) and False Negative (FN) results for the *background* class:

1. False Negative ($Prediction \neq 0$ and $GT = 0$): In this case, the predicted label is simply changed back to '0' (i.e. *background* class).

2. False Positive ($GT \neq 0$ and $Prediction = 0$): In this case, the predicted label is changed from the *background* class to the most likely *other* class. As illustrated in Figure 2, for each FP pixel the non-zero label that appears most frequently in the 3×3 grid around it is selected as the new label. If the grid contains only *background* pixels (i.e. class '0'), then the grid is expanded by one pixel (i.e. 5×5 grid) and the process is repeated until a at least a non-zero label is found. As will be shown, this simple process works well in our context.

As will be shown in the Experimental Results (Section 4), while the baseline semantic segmentation results are already good, this process delivers some additional improvements.

3.4 Semantic segmentation results

The model trained in Section 3.2 is tested using individual orthophotos generated by the process described in [4], which may not necessarily cover entirely a given roof panel of interest or may overlap. To obtain one single orthophoto covering the overall panel with a unified semantic segmentation result, the labels of individual orthophotos must be merged. For this, for each pixel, the label that appears most frequently among the unmerged orthophotos is selected as the final label. In cases when two (or more) classes have the same frequencies for all unmerged orthophotos, the final label will be selected in the following order of priority: *other*, *leadwork*, *slate*. For example, if *slate* and *other* appear the same time for one pixel, then *other* would be chosen as the final label, because we observed that the *other* objects always lay above the *slate* surface.

4 Experimental results

In this section, evaluation metrics for semantic segmentation are first introduced. Then all the experimental results are reported. First, different models are trained using the Duff House training dataset and tested using the Duff House and St Mary's Church testing datasets. The best model is selected by comparing these results, and it is finally re-trained and tested using the combined dataset.

All the training and testing work is completed in the Google Colab Pro environment, with NVIDIA A100 GPU 40 GB.

4.1 Evaluation metrics

The segmentation results for each class are evaluated using two parameters: *Intersection over Union (IoU)*: computed by contrasting the Prediction and Ground Truth segmentations; and *Accuracy (Acc)*: calculated by dividing the sum of the True Positive pixels by the sum of the True Positive pixels and False Positive pixels. To compare the testing performance between different models, the following overall evaluation metrics are evaluated:

- *aAcc*: the Accuracy of all pixels, evaluating the classification accuracy.
- *mIoU*: the mean *IoU* of all classes; *mIoU* is an important indicator to measure the accuracy of overall semantic segmentation.
- *mAcc*: the mean *Acc* of all classes, evaluating the overall performance in pixel classification.

In the following, we report results for different models in the form of tables and confusion matrices. In the tables, we report for each model: the *aAcc*, *mIoU* and *mAcc* for both overall and overall (excl. *background*), and then *IoU* and *Acc* for each of the four classes. For the confusion

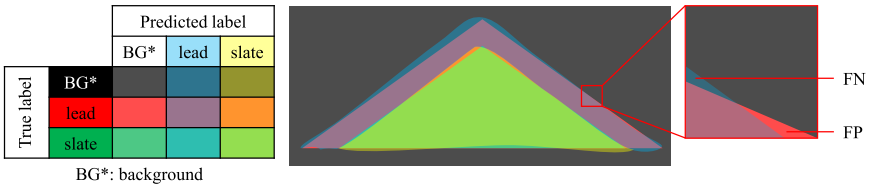


Figure 1. Colour coded confusion matrix (left), semantic segmentation model output (middle), illustrations of FP/FN from the perspective of *background* class(right).

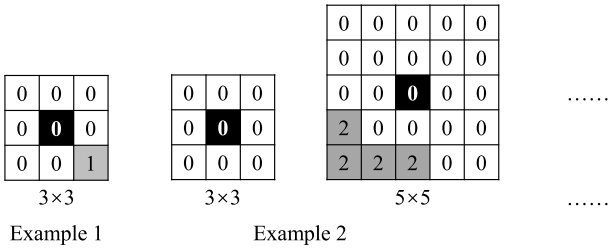


Figure 2. Illustration of the correction of FP for the *background* class. Left: Example 1 where the label is corrected to '1' after one step; Right: Example 2 where the label is corrected to '2' after two steps.

matrices, we report both absolute (in pixel counts) and relative (in percentages) confusion matrices.

4.2 Background label correction

Regardless of the model employed, the *Background model correction* step described in Section 3.3 can be applied to correct FP and FN for the *background* class. We thus demonstrate the benefits of this correction using one model (which we will see later performs well): PointRend (CEL+FL).

Table 2 and Figure 3 show the results obtained when training and testing this model on Duff House dataset, before and after applying the *background* label correction results.

While the baseline performance is already quite good (all metrics > 85 and most of them are > 95), the additional corrective step improves performance for all classes, in particular the *leadwork* and *other* classes,which had the lowest performance without this correction. The higher increases in mIoU (+1.81 to 94.64) and mAcc (+1.19 to 97.24) also indicate a reduced difference in performance among the different classes. Although anticipated, this improvement is welcome, because the *background* regions are often next to *leadwork* regions. And so, any correction of *background* class would most likely benefit the *leadwork* class. Nonetheless, the results demonstrate the good performance of the proposed *background* label correction method.

4.3 Initial models with Duff House dataset

All models are first trained using Duff House training dataset only, and tested with the Duff House testing dataset. Table 3 presents the performance of the different models. Generally, all models already show good performance: the evaluation metrics of overall performance are all > 90, most of them are > 95 and even nearly 100. Generally, the target class, *slate*, is segmented satisfactorily. Errors mostly come from the classes *leadwork* and *other*.

By comparing the first 4 rows, PointRend stands out with the highest values in all evaluation metrics. The variants in the last three rows are then developed based on PointRend, in an attempt to enhance performance with regard to specific challenges with our dataset, namely data class imbalance (the *leadwork* and *other* classes occupy much fewer pixels than *background* and *slate*) and region ambiguity.

Among all the variant models, PointRend(CEL+FL) achieves the best overall performance. Though it is close to the default PointRend model, it increases the values on *other* class both in IoU (+0.89 to 88.52) and Acc (+0.88 to 94.76). In comparison, PointRend(CEL+DL), by adding the Dice Loss function, also improves the performance on *leadwork* class, but with a sacrificial drop on *other* class and accordingly, a decreased overall performance. Therefore, from all the 3 variant models, we conclude that the Focal Loss more successfully improves the results and accounts for data imbalance better than Dice Loss.

4.4 Testing generalisation with St Mary's Church testing dataset

Table 4 reports the evaluation results of all models developed in Section 4.3 on the testing dataset of St Mary's Church. This enables an assessment of the model's generalisation, since no data from St Mary's Church was used to train those models. Table 4 shows a similar pattern as Table 3: all models can segment the *slate* area more accurately than *leadwork* and *other* classes, while the *other* class has the worst performance among all classes. However, compared to the performance on Duff House, there are general decreases in the overall performance of all models, such as a nearly 30 drop in mIoU and nearly 20

Table 2. Testing results before and after background label correction

Result	overall			overall (excl. *background*)			*background*		*slate*		*leadwork*		*other*	
	aAcc	mIoU	mAcc	aAcc	mIoU	mAcc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc
Before	98.97	92.83	96.05	-	-	-	99.26	99.77	98.26	99.0	88.57	92.91	85.26	92.5
After	99.36	94.64	97.24	98.52	92.86	96.32	100	100	98.57	99.3	91.49	94.9	88.52	94.76

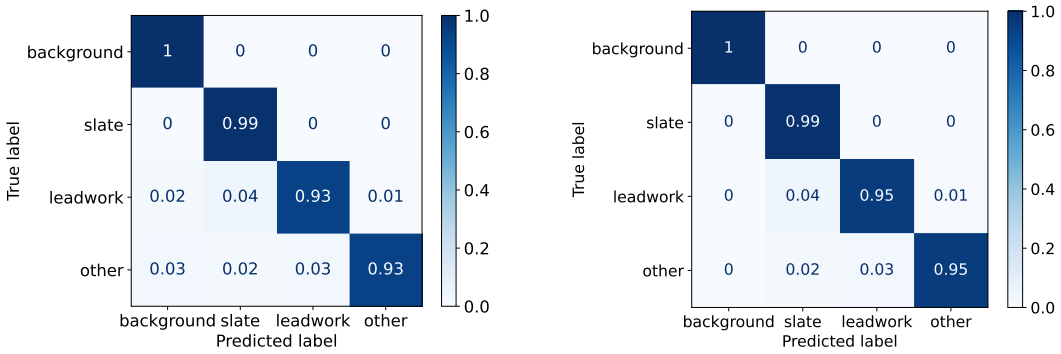


Figure 3. Confusion matrices before (left) and after (right) applying the *background* label correction step.

Table 3. Testing results on Duff House

Model	overall			overall (excl. *background*)			*background*		*slate*		*leadwork*		*other*	
	aAcc	mIoU	mAcc	aAcc	mIoU	mAcc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc
DeepLabV3	99.24	93.54	96.73	98.24	91.39	95.64	100	100	98.37	99.25	89.49	92.48	86.31	**95.18**
PointRend	**99.36**	94.46	97.11	**98.52**	92.61	96.15	100	100	**98.59**	99.3	**91.62**	95.28	87.63	93.88
Segformer	99.17	93.43	96.63	98.08	91.23	95.51	100	100	97.99	98.97	89.6	94.05	86.11	93.51
Segmenter	99.03	92.06	94.73	97.75	89.41	92.97	100	100	97.81	99.31	86.9	92.22	83.52	87.39
PointRend(CEL+DL)	99.03	90.78	93.51	97.75	87.71	91.34	100	100	98.1	**99.45**	88.4	**96.16**	76.63	78.42
PointRend(CEL+FL)	**99.36**	**94.64**	**97.24**	**98.52**	**92.86**	**96.32**	100	100	98.57	99.3	91.49	94.9	**88.52**	94.76
PointRend(DL+FL)	99.33	94.46	96.91	98.45	92.61	95.88	100	100	98.45	99.35	91.48	94.37	87.9	93.93

drop in mAcc. When comparing models in different rows, performance values for the *leadwork* and *other* classes are lower and more spread out than the *other* class and the same classes in Table 3. In general, there is no prominent model that stands out in all metrics specifically for St Mary's Church.

Therefore, even though aAcc remains close to 100% and the general performance is acceptable, the generalisation ability of the models is limited. Put another way, unsurprisingly the models, when trained using data from only one building (Duff House), work but not sufficiently well for other buildings. Therefore, more data, especially with diverse features of *leadwork* and *other* classes, should be used for training.

4.5 Models trained with combined dataset

Based on the results above, the best performing model, PointRend (CEL+FL), is retrained using the the combined training dataset, and tested the combined testing dataset.

The overall performance, reported in Table 5, is almost at the same level as that of Table 3, whose performance was already very high. Looking at individual classes, a slight improvement is achieved for the *slate* class with IoU (+0.04 to 98.61) and Acc (+0.25 to 99.55). However, some reductions in performance are observed for the *leadwork* and *other* classes. But, importantly, compared to the results in Table 4, all metrics show significant improvements. This implies that the new model has achieved a greater level of generalisability (it performs well on testing data from both Duff House and St Mary's Church) without significant drop in overall performance. Naturally, this does not mean the new model will work in all cases of slated roofs; much more diverse data would need to be collected for that. But, the selected model performs satisfactorily.

4.6 Visualise semantic segmentation results

After merging the result labels using the strategy in Section 3.4, the resulting confusion matrix is reported in Figure 4. It shows that the *slate* class segmentation accuracy is still high, but the confusion between *leadwork* and *other* is not insignificant.

This is further illustrated with three example roof panels in Figure 5 and Figure 6. The segmentation result of Panel A is nearly flawless. In Panel B, though the situation is more complex (containing *other* pixels), the slated area of this orthophoto is generally segmented satisfactorily. However, there are still some pixel misclassifications specifically caused by the *other* class, with confusions observed particularly between *slate* and *other* at the bottom of this panel. In the result of Panel C, there are noise pixels

Table 4. Testing results on St.Mary's Church by models trained using Duff House training dataset

Model	overall			overall (excl. *background*)			*background*		*slate*		*leadwork*		*other*	
	aAcc	mIoU	mAcc	aAcc	mIoU	mAcc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc
DeepLabV3	96.48	62.74	69.51	91.84	50.32	59.35	100	100	96.42	98.56	31.3	37.51	23.23	41.97
PointRend	96.66	65.65	75.62	92.26	54.2	67.49	100	100	96.37	97.9	45.93	75.46	20.3	29.11
Segformer	96.89	**66.72**	76.74	92.79	**55.63**	68.99	100	100	**97.25**	98.14	43.48	65.41	**26.15**	43.41
Segmenter	**97.21**	66.58	74.05	**93.53**	55.44	65.4	100	100	96.91	**99.1**	**53.94**	78.69	15.48	18.42
PointRend(CEL+DL)	96.39	64.17	76.33	91.63	52.23	68.44	100	100	96.12	97.3	43.48	**85.3**	17.1	22.73
PointRend(CEL+FL)	94.79	64.31	**80.45**	87.92	52.41	**73.93**	100	100	92.94	93.89	41.93	78.25	22.35	**49.66**
PointRend(DL+FL)	94.6	62.5	74.24	87.48	50	65.66	100	100	93.11	94.52	40.63	60.9	16.27	41.55

Table 5. Testing results on combined dataset

Model	overall			overall (excl. *background*)			*background*		*slate*		*leadwork*		*other*	
	aAcc	mIoU	mAcc	aAcc	mIoU	mAcc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc
PointRend(CEL+FL)	99.0	88.91	92.77	97.68	85.22	90.35	100	100	98.61	99.55	85.93	91.99	71.11	79.52

at the top of the roof panel, the *GT* of which is *slate* but predicted as *other*. The reason for this mistake possibly lies in the biological growth on the slate surface, which may still confuse the deep learning model despite some of the training data containing it. This issue may nonetheless be addressed through a more extensive training dataset.

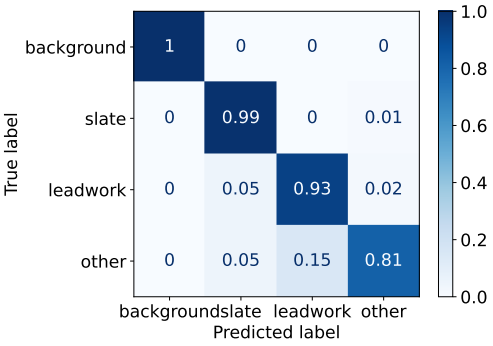


Figure 4. Confusion matrices of results after applying merging strategy

5 Conclusions

Different deep learning models for semantic segmentation are developed and compared using a dataset composed of data coming from two traditional buildings: Duff House and St Mary's Church. PointRend added with Focal Loss (PointRend(CEL+FL)), trained by the combined dataset is chosen as the most suitable when considering both datasets jointly. All of its evaluation matrices, except the ones in *other* class, are all > 85%.

The performance of all models is enhanced thanks to an extra *background* label correction steps: by eliminating the confusion between *background* and other classes (especially *leadwork* and *other*, which are usually the surrounding area of *slate*), the *background* accuracy is corrected to 100%, and the accuracies of other classes are shown to also increase.

However, the confusion matrix in Figure 4 shows that the confusion between *leadwork* and *other* is still significant, affecting the accuracy of *other*. This is possibly because that the *other* class includes many kinds of objects. While enhancing the model robustness may be achieved with more data, it must be highlighted again that our main focus is the effective segmentation of the *slate* and *leadwork* classes.

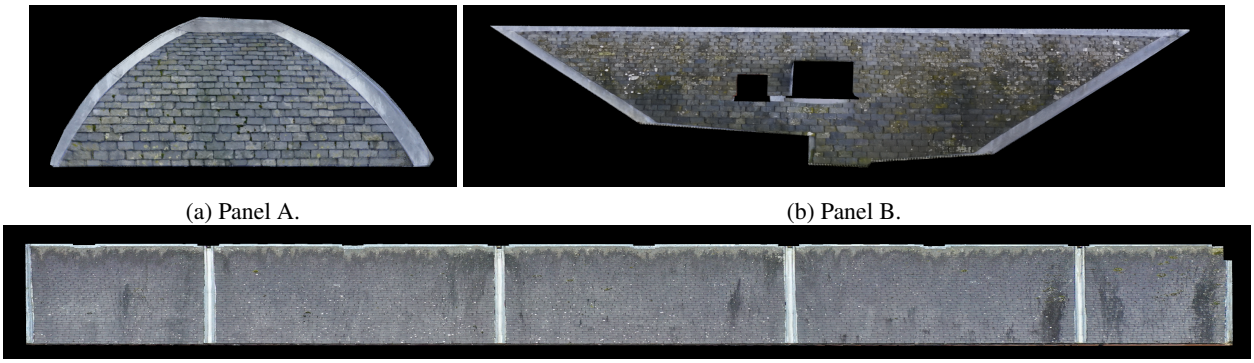
The proposed method focuses on the traditional building roofs, but the methodologies developed are equally applicable to more modern roofs with a slate or tile construction.

It should be noted that the 'best' model is just marginally better than the other ones, with all of them perform reasonably well. The difference of testing results on Duff House and St. Mary's Church indicates the risk of over-fitting, which can be addressed by getting more data involved. With more data collected and used for training, the strategy of selecting the best and most robust model can be improved. Future work can thus first look at collecting more building roof data in order to further validate and improve the orthophoto generation pipeline, and train semantic segmentation models with greater generalisability. These data shall include roofs with various forms and shapes, including slate laying methods and different components around the slated areas, and in various conditions, containing different levels of deteriorations.

Future work should also look at the next step of our proposed overall pipeline, starting with the detection of defects in the slated areas in particular, and the leadwork areas as well. Due to the fact that all the created orthophotos have the same orientation and uniform scale, we anticipate that this should ease the development of further machine learning models.

6 Acknowledgements

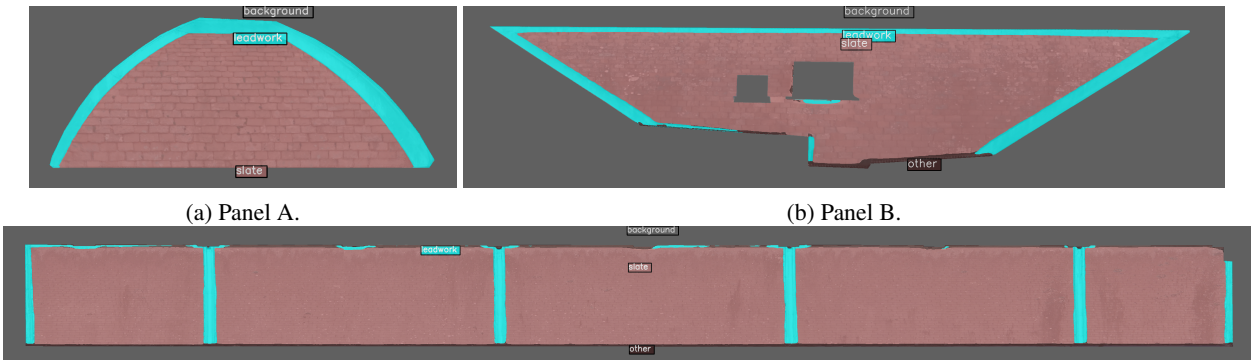
This paper was made possible thanks to research funding from Historic Environment Scotland (HES) and the Engineering and Physical Sciences Research Council (EPSRC) [grant reference EP/W524384/1]. The views and



(a) Panel A. (b) Panel B.

(c) Panel C.

Figure 5. Orthophotos of typical panels.



(a) Panel A. (b) Panel B.

(c) Panel C.

Figure 6. Semantic segmentation results of orthophotos of the typical panels shown in Figure 5

opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of HES and EPSRC. The authors would also like to acknowledge the HES Digital Documentation and Innovation team, Stirling City Heritage Trust for providing us with the data used in the experiments reported in this paper.

References

[1] Roger Curtis and Jessica Hunnisett Snow. Short guide-climate change adaptation for traditional buildings. 2016. URL `https://www.historicenvironment.scot/archives-and-research/publications/publication/?publicationId=a0138f5b-c173-4e09-818f-a7ac00ad04fb`.

[2] J Walsh. Predicting the service life of natural roofing slates in a scottish environment. In *9th international conference on durability of building materials and components. Brisbane: In House Publishing*, 2002. URL `https://inspectapedia.com/roof/Scottish-Roof-Life-Walsh.pdf`.

[3] Min Deng, Carol C Menassa, and Vineet R Kamat. From bim to digital twins: A systematic review of the evolution of intelligent building representations in the aec-fm industry. *Journal of Information Technology in Construction*, 26, 2021. doi:10.36680/j.itcon.2021.005.

[4] Jiajun Li, Frédéric Bosché, Chris Xiaoxuan Lu, and Lyn Wilson. Occlusion-free orthophoto generation for building roofs using uav photogrammetric reconstruction and digital twin data. In *40th International Symposium on Automation and Robotics in Construction*, pages 371–378, 2023. doi:10.22260/ISARC2023/0051.

[5] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew. A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, 7:87–93, 2018. doi:10.1007/s13735-017-0141-z.

[6] Shijie Hao, Yuan Zhou, and Yanrong Guo. A brief survey on semantic segmentation with deep learning. *Neurocomputing*, 406:302–321, 2020. ISSN 0925-2312. doi:10.1016/j.neucom.2019.11.118.

[7] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and Jose Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. *arXiv*, 2017. doi:10.48550/arXiv.1704.06857.

[8] Jong Won Ma, Thomas Czerniawski, and Fernanda Leite. Semantic segmentation of point clouds of building interiors with deep learning: Augmenting training datasets with synthetic bim-based point clouds. *Automation in Construction*, 113:103144, 2020. ISSN 0926-5805. doi:10.1016/j.autcon.2020.103144.

[9] Liu Yang and Hubo Cai. Cost-efficient image semantic segmentation for indoor scene understanding using weakly supervised learning and bim. *Journal of Computing in Civil Engineering*, 37(2):04022062, 2023. doi:10.1061/JCCEE5.CPENG-5065.

[10] Biswajeet Pradhan Abolfazl Abdollahi and Abdullah M. Alamri. An ensemble architecture of deep convolutional segnet and unet networks for building semantic segmentation from high-resolution aerial images. *Geocarto International*, 37(12):3355–3370, 2022. doi:10.1080/10106049.2020.1856199.

[11] Chao Yin, Boyu Wang, Vincent J.L. Gan, Mingzhu Wang, and Jack C.P. Cheng. Automated semantic segmentation of industrial point clouds using respointnet++. *Automation in Construction*, 130:103874, 2021. ISSN 0926-5805. doi:10.1016/j.autcon.2021.103874.

[12] Hanyu Shi, Guosheng Lin, Hao Wang, Tzu-Yi Hung, and Zhenhua Wang. Spsequencenet: Semantic segmentation network on 4d point clouds. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4573–4582, 2020. doi:10.1109/CVPR42600.2020.00463.

[13] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv*, 2017. doi:10.48550/arXiv.1706.05587.

[14] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9796–9805, 2020. doi:10.1109/CVPR42600.2020.00982.

[15] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7242–7252, 2021. doi:10.1109/ICCV48922.2021.00717.

[16] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12077–12090. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf`.

[17] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis E. H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 538–547, 2021. doi:10.1109/ICCV48922.2021.00060.

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. doi:10.1109/ICCV.2017.324.

[19] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, 2016. doi:10.1109/3DV.2016.79.

[20] Michael Yeung, Evis Sala, Carola-Bibiane Schönlieb, and Leonardo Rundo. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*, 95:102026, 2022. ISSN 0895-6111. doi:10.1016/j.compmedimag.2021.102026.